

REMARKS

Claims 1-20 are pending in the application. Please amend Claims 1 and 10. Please add Claims 21-22. The pending claims are recited below for convenience.

1. (Currently amended) A computer-implemented method of determining content type of contents of a subject Web page, comprising the steps of:
 - providing a predefined set of potential content types, content types being exclusive of indicating formal language of the content;
 - for each potential content type, preparing a distinguishing series of tests, wherein the distinguishing series of tests includes at least one binary tests, at least one non-binary tests and at least one test: (a) examining syntax or grammar; or (b) examining page format or style other than position of data or a keyword in the subject Web page;
 - for each potential content type, running the distinguishing series of tests having test results which enable quantitative evaluation of at least some contents of the subject Web page being of the potential content type;
 - mathematically combining the test results; and
 - based on the combined test results, assigning a respective probability, for each potential content type, that some contents of that type exists on the subject Web page, and indicating content type, said indicating being exclusive of indicating language in which content is written.
2. (Original) A method as claimed in Claim 1 wherein the set of potential content types include any combination of organization description, organization history, organization mission, organization products/services, organization members, organization contact information, management team information, job opportunities, press releases, calendar of events/activities, biographical data, articles/news with information about people, articles/news with information about organizations and employee roster.

3. (Original) A method as claimed in Claim 1 wherein the step of combining includes producing a respective confidence level for each potential content type, that at least some content of the subject Web page is of the potential content type.
4. (Original) A method as claimed in Claim 1 wherein the step of combining the test results includes using a Bayesian network.
5. (Original) A method as claimed in Claim 4 further comprising the step of training the Bayesian network using a training set of Web pages with respective known content types such that statistics on the test results are collected on the training set of Web pages.
6. (Previously presented) A method as claimed in Claim 1 wherein the predefined set includes a potential content type of press release and the distinguishing series of tests further includes at least one of:
 - (i) determining whether a predefined piece of data or keyword appears in the subject Web page;
 - (ii) examining text properties; and
 - (iii) determining whether the predefined piece of data or keyword appears in URLs in the subject Web page.
7. (Previously presented) The method as claimed in Claim 1 wherein the distinguishing series of tests further includes at least one of:
 - (i) determining whether a predefined piece of data or keyword appears in the subject Web page;
 - (ii) examining text properties; and
 - (iii) determining whether the predefined piece of data or keyword appears in URLs in the subject Web page.

8. (Original) A method as claimed in Claim 1 further comprising the step of storing indications of the assigned probabilities of each potential content type per respective Web page.
9. (Original) A database formed by the method of Claim 8, said database containing indications of Web pages and corresponding content types determined to be found on respective Web pages.
10. (Currently amended) Apparatus for determining content type of contents of a subject Web page, comprising:
 - a predefined set of potential content types, each potential content type being exclusive of indicating formal language of the content and associated with a respective distinguishing series of tests, wherein the distinguishing series of tests includes at least one binary tests, at least one non-binary tests and at least one test: (a) examining syntax or grammar; or (b) examining page format or style order other than position of data or a keyword in the subject Web page;
 - a test module utilizing the predefined set, the test module employing the distinguishing series of tests as a plurality of processor-executed tests having test results which enable, for each potential content type, quantitative evaluation of at least some contents of the subject Web page being of the potential content type, for each potential content type, the test module (i) running the respective distinguishing series of tests, (ii) combining the test results and (iii) for each potential content type, assigning a respective probability that at least some contents of that type exists on the subject Web page being of the potential content type, and indicating content type exclusive of indicating language in which content is written.
11. (Original) Apparatus as claimed in Claim 10 wherein the set of potential content types include any combination of contact information, press release, company description, employee list, other.

12. (Original) Apparatus as claimed in Claim 10 wherein the test module produces a respective confidence level for each potential content type, that at least some content of the subject Web page is of the potential content type.
13. (Original) Apparatus as claimed in Claim 10 wherein the test module combines the test results using a Bayesian network.
14. (Original) Apparatus as claimed in Claim 13 further comprising a training member for training the Bayesian network using a training set of Web pages with respective known content types, such that statistics on the test results are collected on the training set of Web pages.
15. (Original) Apparatus as claimed in Claim 10 wherein the predefined set includes a potential content type of at least one of organization description, organization history, organization mission, organization products/services, organization members, organization contact information, management team information, job opportunities, press releases, calendar of events/activities, biographical data, articles/news with information about people, articles/news with information about organizations and employee roster.
16. (Previously presented) Apparatus as claimed in Claim 15 wherein the processor-executed tests include at least one of:
 - (i) determining whether a predefined piece of data or keyword appears in the subject Web page;
 - (ii) examining text properties; and
 - (iii) determining whether the predefined piece of data or keyword appears in URLs in the subject Web page.
17. (Previously presented) Apparatus as claimed in Claim 10 wherein the processor-executed tests include any of:

(i) determining whether a predefined piece of data or keyword appears in the subject Web page;

(ii) examining text properties; and

(iii) determining whether the predefined piece of data or keyword appears in URLs in the subject Web page.

18. (Original) Apparatus as claimed in Claim 10 further comprising storage means for receiving and storing indications of the assigned probabilities of each content type per Web page as determined by the test module, such that the storage means provides a cross reference between a Web page and respective content types of contents found on that Web page.

19. (Previously presented) A method as claimed in Claim 1 wherein the at least one binary test and the at least one non-binary tests include one or more of the following tests:

- i) whether the subject Web page contains a press release;
- ii) whether the subject Web page has a title;
- iii) whether the subject Web page has a copyright statement;
- iv) whether the subject Web page has a navigation map;
- v) whether the subject Web page has a line with a keyword followed by at least another keyword within the next 10, 20, 30 or 40 lines;
- vii) whether a first sentence of a first paragraph of the subject Web page has a date;
- viii) whether the first sentence of the first paragraph of the subject Web page is preceded by a header line;
- ix) whether the first sentence of the first paragraph of the subject Web page contains the keyword or a form of the keyword;
- xi) whether the subject Web page contains a text line starting with the keyword; and

xii) a calculation of a percentage of header lines, the average sentence length, number of different domains, number of lines that contain the keyword or number of phrases that contain the keyword.

20. (Previously presented) Apparatus as claimed in Claim 10 wherein the at least one binary test and the at least one non-binary tests include one or more of the following tests:

- i) whether the subject Web page contains a press release;
- ii) whether the subject Web page has a title;
- iii) whether the subject Web page has a copyright statement;
- iv) whether the subject Web page has a navigation map;
- v) whether the subject Web page has a line with a keyword followed by at least another keyword within the next 10, 20, 30 or 40 lines;
- vii) whether a first sentence of a first paragraph of the subject Web page has a date;
- viii) whether the first sentence of the first paragraph of the subject Web page is preceded by a header line;
- ix) whether the first sentence of the first paragraph of the subject Web page contains the keyword or a form of the keyword;
- xi) whether the subject Web page contains a text line starting with the keyword; and
- xii) calculation of a percentage of header lines, the average sentence length, number of different domains, number of lines that contain the keyword or number of phrases that contain the keyword.

21. (New) The method of Claim 1, wherein the at least one test examining syntax or grammar includes at least one test measuring one of the following: number of passive sentences, number of sentences without a verb, and percentage of verbs in past tense.

22. (New) The apparatus of Claim 10, wherein the at least one test examining syntax or grammar includes at least one test measuring one of the following: number of passive sentences, number of sentences without a verb, and percentage of verbs in past tense.

Remarks

Amendment to Claims

Claims 1 and 10 have been amended to recite “content type being exclusive of indicating formal language of the content” and “indicating content type being exclusive of indicating language in which content is written” as shown above. Support for amendment can be found generally in the specification as originally filed. No new matter is introduced.

New Claims

Claims 21 and 22 are being added. Support for the claims can be found at, for example, page 12, lines 1-2 of the specification as originally filed.

No new matter is added.

One-Month Extension

This Amendment is submitted responsive to the Final Action mailed on October 16, 2006 with one-month extension, not two. Applicants had submitted a Reply within two months of the mailing date of the Final Action. Subsequently, the examiner issued an Advisory Action after the end of the three-month period, on January 18, 2007. According to § 706.06(f) of the MPEP, if the examiner does not mail an advisory action until after the end of the 3-month period, the shortened statutory period will expire on the date the examiner mails the advisory action and any extension of time fee would be calculated from the mailing date of the advisory action. As such, this Amendment, being mailed on February 20, 2007, is being properly filed with one-month extension because February 18, 2007 falls on Sunday and February 19 is a Federal holiday.

35 U.S.C. § 103(a) Rejection of Claims 1-3, 6-9, 10-12 and 15-20

The present invention is directed to computer-implemented methods and apparatus for determining the content type of a subject Web page. Particularly, in base Claims 1 and 10, the invention includes a distinguishing series of tests. In addition to non-binary tests, the distinguishing series of tests include at least one test: (a) examining syntax or grammar; or (b) examining page format or style other than the position of data or a keyword in the subject Web page. Examining syntax or grammar includes, for example, the number of passive sentences,

number of sentences without a verb, percentage of verbs in past tense, number of fonts used, and/or existence of certain characters in determining the content type of the subject Web page. (See page 7, lines 8-11 and page 19, lines 4-12, Specification.)

Claims 1-3, 6-9, 10-12 and 15-20 have been rejected under 35 U.S.C. § 103(a) as being unpatentable over Russell-Falla et al. (U.S. Patent No. 6,675,162) (hereinafter “Russell-Falla”) in view of Chakrabarti et al. (U.S. Patent No. 6,389,436) (hereinafter “Chakrabarti”) in further view of van den Akker (U.S. Patent No. 6,415,250) (hereinafter “Akker”).

Maintaining the rejection previously issued in the Final Action mailed on October 16, 2006, the examiner stated in the Advisory Action mailed on January 18, 2007 that the arguments for patentability of the claims submitted by Applicants were unpersuasive. The examiner disagreed with Applicants that the cited references as a whole teach away from each other. The examiner is understood to state that Claims 1-3, 6-9, 10-12 and 15-20 are obvious in view of the cited references because:

- (1) Chakrabarti teaches a non-binary test that utilizes the hyperlinks in documents to help classify the document based on the classification of the documents that the hyperlinks access within a certain radius;
- (2) Having found the appropriate documents in the assigned radius, a text-based classifier was then utilized to help assign a probability vector to each document; and
- (3) the examiner believes that the probabilistic analysis of the text of an incoming document shown in Akker would provide the same benefit to Chakrabarti that it would have provided to Russell-Falla.

(See page 2 of the Advisory Action.)

In the Final Action mailed on October 16, 2006, the examiner’s reason to reject Claims 1-3, 6-9, 10-12 and 15-20 under 35 U.S.C. § 103 was that Akker cures a deficiency by Russell-Falla in view of Chakrabarti. (See paragraph 2, page 4 of the Final Action.) The examiner stated that while Russell-Falla does not specifically teach the limitation that at least one test was examining syntax or grammar, Akker teaches a test for classifying an incoming text, “wherein

the test includes probabilistic analysis of the inputted text which reflects morphological characteristics of natural language..., wherein the tests examine the syntax and grammar of the incoming text." (*See id.*) The examiner further stated that "it would have obvious to one ordinary skill in the art at the time of the invention for one of the distinguishing series of tests of Russell-Falla to have analyzed to syntax or grammar of the text". (*See id.*)

The combined teachings of Russell-Falla and Akker do not imply or suggest the present invention

It is Applicants' position that one skilled in the art would not be motivated to modify the teachings of Russell-Falla incorporating "the probabilistic analysis of the inputted text" of Akker, "which reflects morphological characteristics of natural language" to produce the present invention. And in the alternative, if combined, the combination falls short of Applicants' claimed invention. Obviousness under 35 U.S.C. § 103 can only be established by combining or modifying the teachings of the prior art to produce the claimed invention where there is some teaching, suggestion, or motivation to do so. *In re Kahn*, 441 F.3d 977, 986, 78 USPQ2d 1329, 1335 (Fed. Cir. 2006)

Russell-Falla is directed to the methods for scanning and analyzing various kinds of digital information content. One of the listed objectives of the invention by Russell-Falla is to enable parents or guardians to exercise some control over the web page content displayed to their children. (*See col. 2, lines 26-29 of Russell-Falla.*) Furthermore, the invention by Russell-Falla is useful for a variety of applications, including, to blocking digital content, especially world-wide web pages, from being displayed when the content is unsuitable or potentially harmful to the user, or for any other reason that one might want to identify particular web pages based on the content. (*See col. 2, lines 43-49 of Russell-Falla.*) Russell-Falla includes a step of identifying and analyzing the web page natural language content relative to a predetermined database of words – or more broadly regular expressions – to form a rating. (*See col. 2, lines 52-59; col. 3, lines 24-27; and col. 4, lines 59-61 of Russell-Falla.*) The database includes a list of unique words (*i.e.* "breast") and phrases (regular expressions) that are useful in discriminating a specific category of information such as pornography. (*See col. 2, lines 61-66; and col. 7, lines 17-24 of Russell-Falla.*)

Akker is directed to an automatic language identification system for determining the source of language. (See Abstract and col. 3, lines 26-28 of Akker.) As the examiner stated in the Advisory Action, Akker discloses “the probabilistic analysis of the inputted text which reflects morphological characteristics of natural language”. (See col. 3, lines 4-9 of Akker.) According to Akker, the “probabilistic analysis” of the unknown text identifies the language (*i.e.* English, German, Dutch or French) in which a text is written based upon predetermined word portions (*i.e.* infix, prefix or suffix of the word) containing morphological features of the language, by the way word forms are produced from word roots. (See Figs. 2A-2C and col. 8, line 60 – col. 9, line 3 of Akker.)

In the present invention as claimed, the “probabilistic analysis” is in the context of determining content type of contents of a subject Web page, content type being “exclusive of indicating language in which the content is written.” However, the “probabilistic analysis” in Akker fails to provide any information as to determining such content type of a subject Web page,”. As discussed earlier, the invention in Akker identifies the language in which a text is written based upon predetermined word portions containing morphological features of the language. Exemplifying a linguistic tool of the “probabilistic analysis”, Akker discloses a suffix extractor 404 that extracts the suffix of each of the parsed words. (See Fig. 4 and col. 11, line 65 – col. 12, line 26 of Akker.) Subsequent to extraction of suffixes, the “probabilistic analysis” tallies the frequency of each suffix appeared using the suffix frequency list generator 406. (See Fig. 4 and col. 12, lines 33-54 of Akker.) However, identifying one or more suffixes and counting frequency of each of the suffixes appearing in the subject Web page cannot provide any insight on the content type of the subject Web page, exclusive of language in which content was written. Finding out that a suffix, for example, “-er” or “-ly”, appears eight times cannot be insightful in determining whether the Web page is related to sports or finance and cannot be insightful in determining age inappropriate material for Russell-Falla purposes.

Furthermore, nowhere in Akker is there a teaching, disclosure or suggestion that these predetermined word portions are related to a content type matter except to identifying a language or that any linguistic tool for determining a content matter. However, such identification of the language is also extraneous in determining content type of contents of a subject Web page as in

the claimed present invention. That is, finding out that the subject Web page is written in German does not impart that the Web page is related to pornography.

As such, Akker does not teach examining syntax or grammar in the context of determining content type of contents of a subject Web page as in the present invention. Contrary to the examiner's view that Akker provides a benefit to Russell-Falla, the linguistic tools in Akker, in fact, provides no advantage in determining content types of contents of a subject Web page. While Akker may arguably disclose examining syntax or grammar, one skilled in the art cannot be motivated to combine the teachings of Russell-Falla and Akker to arrive at the present invention because an identification of the language written in the Web page using Akker is of no use in determining content types of the Web page. Similarly, even if combined, Russell-Falla modified by Akker does not provide the present invention "indicating of content type...exclusive of indicating language in which content is written" as recited in base Claims 1 and 10 as now amended.

One of ordinary skill in the art is not motivated to combine the teachings of Chakrabarti and Akker in view of Russell-Falla

The disclosure of Chakrabarti is related to computer-implemented classifiers, and, in particular, to a hypertext classifier that classifies documents that contain hyperlinks. (See col. 1, lines 6-8 and col. 5, lines 46-53 of Chakrabarti.) Chakrabarti states the following:

A text-based classifier classifies the documents based only on the text contained in the documents. However, *documents on the Web typically contain hyperlinks*. These hyperlinks are ignored by text-based classifiers, although the hyperlinks contain useful information for classification...*There is a need in the art for an improved classifier that can classify documents containing hyperlinks....* Unlike conventional systems, the hypertext classifier 110 of the present invention exploits *topic information* implicitly present in hyperlink structure.

(See col. 2, lines 29-34 col. 3, lines 63-64 and col. 7 lines 5-7 of Chakrabarti; italics added)

As the title of Chakrabarti, which is ENHANCED HYPertext CATEGORIZATION USING HYPERLINKS, indicates, the system of Chakrabarti uses the hyperlinks such as <http://www.research.att.com/lewis> or <http://medir.ohsu.edu/pub/ohsumed> to classify a document

into categories such as “news, entertainment, sports, business or theater.” (See col. 6, lines 62-64 of Chakrabarti.)

As addressed above with respect to Russell-Falla, one skilled in the art would not be motivated to combine the teachings of Chakrabarti and Akker in view of Russell-Falla to arrive at Applicants’ invention. Again, Applicants point out that combining Akker and Chakrabarti to incorporate the “probabilistic analysis” of the text in a hyperlink may identify the language in which the hyperlink text is written based upon predetermined word portions containing morphological features of the language. The morphological analysis taught in Akker, however, does not impart that the hyperlink and the Web page is related to a particular category such as “news, entertainment, sports, business or theater.” Furthermore, because hyperlinks often contain incomplete words, a string of letters or acronyms, there may not be sufficient frequencies of predetermined word portions to extract a critical mass of data. As such, one skilled in the art cannot be motivated to combine the teachings of Chakrabarti and Akker for use with Russell-Falla to arrive at the present invention. Similarly, even if combined, the combination of Akker, Chakrabarti and Russell-Falla does not provide the “content type ...exclusive of indicating language in which content is written” as recited in base Claims 1 and 10 as now amended.

In view of foregoing, Applicants respectfully request the § 103(a) rejection of base Claims 1 and 10 be withdrawn. Furthermore, as Claims 2, 6-9, 11-12 and 15-20 depend from Claim 1 or 10, these claims are allowable for the same reasons discussed above.

35 U.S.C. § 103(a) Rejection of Claims 4, 5, 13 and 14

Claims 4, 5, 13 and 14 have been rejected under 35 U.S.C. § 103(a) as being unpatentable over Russell-Falla in view of Chakrabarti in view of Akker and in further view of Haug et al. (U.S. Patent No. 6,556,964) (hereinafter “Haug”). Haug is directed to a probabilistic model for determining the meaning of sentences or phrases in medical reports. The Haug model extracts and encodes medical concepts using a Bayesian network.

Claims 4-5, 13 and 14 depend from base Claims 1 or 10. Therefore, Claims 4, 5, 13 and 14 also include the element of the distinguishing series of tests having both binary and non-binary tests and the distinguishing claim term “content type...exclusive of indicating language in

which content is written". As explained above, neither Russell-Falla, Chakrabarti nor Akker teaches, suggests or otherwise makes obvious the distinguishing series of tests having one or more tests involving syntax, grammar, page style or the content type being "exclusive of indicating language in which content is written" of Claims 4, 5, 13 and 14. Furthermore, Haug does not cure this deficiency to make Claims 4, 5, 13 and 14 obvious. Therefore, Applicants respectfully request the § 103(a) rejection of Claims 4, 5, 13 and 14 be withdrawn.

Patentability of Claims 21 and 22

Claims 21 and 22 are new and each depend from base Claim 1 and 10, respectively. As such, , Claims 21 and 22 also include the element of the distinguishing series of tests having both binary and non-binary tests and the distinguishing claim term "content type...exclusive of indicating language in which content is written". Accordingly, Claims 21 and 22 are allowable for the same reasons discussed above.

CONCLUSION

In view of the above amendment and remarks, it is believed that all claims (claims 1-22) are in condition for allowance, and it is respectfully requested that the application be passed to issue. If the Examiner feels that a telephone conference would expedite prosecution of this case, the Examiner is invited to call the undersigned.

Respectfully submitted,

HAMILTON, BROOK, SMITH & REYNOLDS, P.C.

By 
H. Joon Chung
Registration No. 52,748
Telephone: (978) 341-0036
Facsimile: (978) 341-0136

Concord, MA 01742-9133

Dated:

2/20/07